TWITTER ARCHIVING USING TWAPPER KEEPER: TECHNICAL AND POLICY CHALLENGES

Brian Kelly
UKOLN,
University of Bath,
Bath, UK

Martin Hawksey
JISC RSC Scotland North & East,
Edinburgh's Telford College,
Edinburgh, UK

John O'Brien
3930 Rolling Hills Drive,
Cumming, GA 30041,
USA

Marieke Guy
UKOLN,
University of Bath,
Bath, UK

Matthew Rowe
Department of Computer Science,
University of Sheffield,
Sheffield, UK

ABSTRACT

Twitter is widely used in a range of different contexts, ranging from informal social communications and marketing purposes through to supporting various professional activities in teaching and learning and research. The growth in Twitter use has led to a recognition of the need to ensure that Twitter posts ('tweets') can be accessed and reused by a variety of third party applications.

This paper describes development work to the Twapper Keeper Twitter archiving service to support use of Twitter in education and research. The reasons for funding developments to an existing commercial service are described and the approaches for addressing the sustainability of such developments are provided. The paper reviews the challenges this work has addressed including the technical challenges in processing large volumes of traffic and the policy issues related, in particular, to ownership and copyright.

The paper concludes by describing the experiences gained in using the service to archive tweets posted during the WWW 2010 conference and summarising plans for further use of the service.

1. ABOUT TWITTER

Twitter has been described as a 'micro-blogging' service. It provides blogging functionality, but the blog posts (often referred to as 'tweets') are restricted to 140 characters. Although this constraint may appear to provide a severe limitation on use of Twitter in an educational and research content, in practice the ease of creating tweets (without the need for the individual to spend time and mental energy in composing their thoughts and perhaps having ideas reviewed by others or checked by an editorial board) has given rise to Twitter being used to support educational and research activities in ways which had not previously being considered. Twitter's popularity has been enhanced by the ability to publish material on a wide range of devices and in particular mobile devices where the 140 character constraint is less of an issue for the small (or virtual) keyboards to be found on such devices.

2. HOW TWITTER IS BEING USED

The growing importance of preservation of Twitter content is illustrated by two examples of existing use to support education and research.

2.1. Supporting Events

Twitter has been used to support a number of high profile events in the UK's higher education community. It has been used by delegates, both physical and virtual, to engage in discussion and disseminate resources. The international ALT-C 2009 conference, which was held over 3 days in September 2009 generated 4.317 Twitter posts from 633 contributors using the #altc2009 hashtag [9]. The JISC's recent annual one-day conference was held in April 2010 generated 2,801 tweets from 479 contributors using the #jisc10 hashtag [10].

UKOLN's annual Institutional Web Management Workshop (IWMW) has made use of networked technologies to support events since an IRC channel was used to support discussions at IWMW 2005. In recent years Twitter has been used; at the IWMW 2009 event the #iwmw2009 hashtag was used with 1,530 tweets posted from 170 contributors [11]. In addition to the identification of a recommended hashtag for the event the organisers set up a dedicated Twitter account to send announcements as well as providing an official commentary of some of the sessions. One of the plenary speakers at the IWMW 2009 also used Twitter in an innovative way, abandoning use of PowerPoint or other presentation tools, instead simply speaking and responding to tweets from the audience (and a remote audience who were following the event's hashtag) which were displayed on a large screen in the auditorium [14].

2.2. Captioning Video

Work on the use of Twitter as a method for captioning videos has been ongoing since 2009 [7]. The core

concept is to convert tweets posted during a live event into a compatible caption file format which then can be replayed with audio or video clips. The development of Twitter based captions has mirrored the increasing use of Twitter to support events and provides a means for delegates to replay archived audio and video recordings with the real-time stream of tweets, in essence allowing users to replay conference sessions augmented with the original backchannel communication. More recently this work has been extended allowing users to generate and play subtitles for on demand television services such as the BBC's iPlayer [2] and political speeches [3].

The software has been developed to use an event hashtag not only to generate subtitles but also to use this resource to allow users to search within the associated media asset. This development opens up the use of Twitter subtitles as a tool to support the increasing popular use of lecture capture in education; that is as well as students being able to replay a captured lecture they can also view the back channel discussions [4].

2.3. Observing Political Debate

Twitter provides real time information about a diverse range of topics, in essence allowing users to be harnessed as social sensors. For instance work by Sakaki [20] has found that Twitter users in Japan could be used as sensors to detect earthquakes by observing their tweets and the location where they were published.

Politics is one of the most discussed topics on Twitter, allowing public sentiment to be gleaned from the analysis of tweets. For instance, work by [1] performed sentiment analysis over a corpus of tweets archived in the run up to the US presidential election of 2008. This allowed public reaction to policy decisions and speeches to be gauged without the need for exhaustive polling.

The archival of tweets discussing politics provides a useful backdated corpus which can be used to explore public reaction and sentiment. Observations made over such data could in turn allow informed future policy decisions to be made.

2.4. Additional Uses

We have shown examples which demonstrate how Twitter is being used within the Higher Education sector. The use of Twitter at events illustrates reasons why tweets should not be regarded as possible value only at the time they were posted as increasingly we might expect to see tweets being analysed after an event in order to inform the evaluation of the event.

Additional reasons why there is a need to ensure that tweets should be made available for reuse include:

- To allow for analysis of Twitter communities e.g. analysis of Twitter spammers [5].
- Analysis of tweets associated with a hashtag used to support sharing and community-building across development programmes such as the JISC Rapid Innovation programe [19] as described at [6].
- Reputation management for both organisations and individuals.
- Personal interests: e.g. to enable a Twitter user to be have an answer to the question "What was I saying when I was young?".

3. WHY THE NEED FOR AN ARCHIVING SERVICE?

Since Twitter provides a search interface to its service it may not be apparent why a third party service is needed to provide an additional archive of tweets.

A key reason which has led to the development of a number of Twitter archiving services is due to limitations of the Twitter search API which provides access only to recently posted tweets. Current documentation from the Twitter Search API states that searches are limited to 1,500 individual tweets from the last 7 days [15]. Consequently as well as not being able to access tweets older than 7 days, the complete timeline for popular 'trending' topics are also not available.

To date Twitter has been designed to facilitate development of third party services around its service, avoiding the need for new features having to be provided by Twitter. Useful additional features which have evolved include statistical analyses, enhanced search capabilities, the ability for end users to manage their collection of Twitter archives and export the data in a variety of formats. As well as such services aimed at end users Twitter archiving services can themselves provide APIs which allow them to be used to provide additional services to developers.

An awareness of the importance of Twitter archiving to the UK's tertiary education community led to the JISC exploring options for a Twitter archiving service.

4. TWITTER ARCHIVING SERVICE OPTIONS

A variety of archiving services for tweets are available. These include WTHashtag (archives tweets for specified hashtags – see http://wthashtag.com/); BackUpMyTweets (used by a Twitter user to provide a backup of their own tweets – see http://backupmytweets.com/) and Twapper Keeper (see http://twapperkeeper.com/).

These services are based on the Twitter APIs. It would be therefore possible to develop a new service for archiving tweets. However, as identified in the JISC 2010 Strategy "*The New JISC Strategy comes amid serious economic recession in the UK*" [8]. Despite such economic concerns, as the JISC Strategy document identifies: "*Cloud computing offers flexibility and, where the business case is done carefully and accurately, can offer considerable savings by avoiding the cost of owning and large computer facilities and the associated running costs*".

Such strategic considerations provided the context in which, instead of commissioning development of a prototype which, if successful, would be expected to

evolve into a national service, it was felt that a more cost-effective development route would be to fund developments of an existing service to ensure that needs of the UK's higher educational sector were addressed.

Following negotiations the JISC agreed to fund developments to Twapper Keeper for a 6 month period from April 2010, with UKOLN, a JISC-funded centre of expertise in digital information management, providing the project management for this work.

The Twapper Keeper blog was used to announce the development work and invite suggestions on developments [17]. The suggestions which were received included:

- Ability to group collections of archives.
- Ability to delete tweets from the Twapper Keeper's archives.
- Ability to opt-out of being archived.
- Provision of APIs to the Twapper Keeper service.
- Access to archives provided in multiple formats (e.g. RSS, Atom and JSON).

There is a need to ensure that the JISC investment in this development work provides a sustainable service. This challenge is nothing new – project-funded work carried out in UK higher educational institutions cannot be guaranteed to result in the delivery of sustainable services. However funding of an external service based outside the UK, while not new, does necessitate that careful attention is taken to not only the development work itself but also the sustainability of the service after the development work is over.

The approaches which are being taken in the development work include:

- **An open approach to developme**nt: to gain buy-in from the user community and ensure developments reflect the communities' needs.
- **Migration of the platform**: to a more stable platform to ensure that the service can cope with the anticipated growth in use and traffic.
- **Open sourcing components**: which would allow the service to be replicated if this was felt to be necessary.
- **Open content for documentation**: through use of Creative Commons licences for the project blog, technical documentation, FAQs, etc.

5. CHALLENGES

Following the gathering of the user requirements for developing the Twapper Keeper service we have prioritised the requirements and developed an implementation plan.

The following technical and policy challenges in implementing the user requests have been identified:

5.1. Technical Issues

Due to the rapid adoption of the Twitter service, the Twitter API and ecosystem continues to evolve. This requires services such as Twapper Keeper to continue to develop and align with the Twitter technical and policy changes. For example, recent changes to the way Twitter recommends tweet data be accessed and consumed from RESTful search service to Streaming APIs has shifted the load of processing tweets to Twapper Keeper forcing the system to take on more burden when trying to store large numbers of tweets.

Another example is alignment with policy requirements for deletion of tweets. Twitter requires third party systems leveraging the API to delete tweets when a user deletes a tweet from Twitter, which requires a delete event to be sent from Twitter to the third party system. However limitations in the current RESTful search / timeline APIs and the Streaming API for tracking keywords results in no deletion events being sent to the Twapper Keeper service. Therefore, to overcome this from a policy perspective the Twapper Keeper service requirements users to inform the service if they want their tweets to be deleted.

A final example of the technical challenges is that changes continue to take place in the Twitter ecosystem, as highlighted by the upcoming plan to drop support for Basic HTTP Authentication with Twitter API REST services and the requirement for services to migrate to use OAuth. The ability of the Twapper Keeper service to manage such changes in accessing data held by the Twitter service enables third party services to avoid the need to make modifications to their service when the Twitter backend access mechanisms are changed.

5.2. Policy Issues

The ability for a user to delete their tweets from the Twapper Keeper archive has been requested as well as users being able to opt-out completely from the service. This request reveals uncertainties regarding the copyright status of Twitter posts and the ways in which third party services should address issues of ownership and related management issues.

The Twitter terms of service state that "*You retain your rights to any Content you submit, post or display on or through the Services*" [16]. This could be interpreted to mean that it would not be possible for tweets to be harvested by others without permission of the owner. However this is clearly not a scalable solution as can be seen by the popularity of Twitter archiving services (including the announcement that the US Library of Congress is to archive tweets [13]). However rather than disregarding concerns of the rights holders Twapper Keeper developments will allow Twitter users to delete their tweets from the Twapper Keeper archive. In addition they will be able to opt out of the Twapper Keeper archive service.

A more challenging request has been to restrict the archiving of a Twitter user's stream of tweets to the owner. Such a requirement could hinder the

development of other user requests (e.g. the ability to archive tweets from a list of Twitter users). It can also be argued that since an individual's tweets can be accessed by using the Twitter search facility users it would be unreasonable to expect that a stream of an individual's tweets should not be archived. From this perspective the issue of, for example, archiving of tweets which might be embarrassing to the poster should be regarded as an educational and new media literacy issue, on par with understanding the risks of sending inappropriate messages to public mailing lists. However we acknowledge that here is a need to address the concerns raised by this request. We are therefore planning to remove the ability to allow open archiving of an individual's tweets; instead users will need to login (via OAuth) and will only be able to create a public archive of their own Twitter stream.

It should be noted that it will still be possible to archive tweets based on keywords. Since the keywords could coincide with a Twitter ID it is possible to find tweets which may refer to an individual. Removing the ability to archive by keywords would undermine the credibility of the service and could result in users migrating to an alternative service. Our approach to this dilemma is to raise awareness of the ways in which an individual's tweets could be archived on the service's FAQ and remind users of possible risks in posting public tweets and the mechanisms for deleting tweets from the Twapper Keeper archive and from Twitter.

5.3. Sustainability Issues

As the Twapper Keeper service continues to grow various issues related to the quality of the service are beginning to arise including:

- Servers are being over-utilised.
- Continuity of backup service is not optimized.
- Users have limited visibility to items that are still being queued for archiving.
- Resource contention on various backend archiving processes needs to be tuned to support the increased number of terms being archived.

In order to address these issues the following actions have been taken:

- A dedicated server (versus a virtual server) has been procured and setup to host the service.
- Additional disk space has been added to provide primary and backup storage on the dedicated server.
- New system monitoring services have been implemented to provide status to system administrators and end users.
- On-going monitoring, refactoring and tuning of archiving algorithms in order to improve the archiving efficiency and effectiveness.

6. EXPERIENCES

6.1. Use of the Twapper Keeper Service

Twapper Keeper was used to archive tweets from the World Wide Web 2010 conference, held in Raleigh North Carolina on 28-30 April 2010 as illustrated in Figure 1. On 7 May 2010 3,616 tweets which used the #www2010 hashtag had been harvested from a total of 909 users.

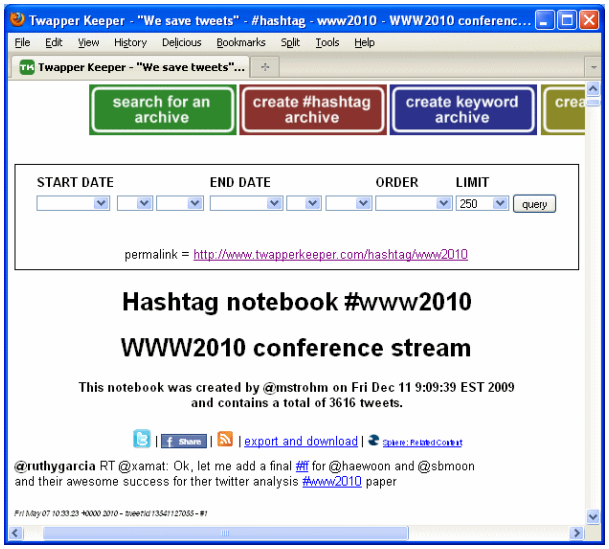

Figure 1: Twapper Keeper interface for the #www2010 archive

During the development work we became aware that tweets were missing from the archive [18]. We discovered that gaps can be introduced during disconnects / reconnects especially if there is a latency in the data transferred between Twitter and Twapper Keeper; if the latency is high, data could be lost. Twapper Keeper now runs a background process that uses the REST /search API to check to see if we have missed any tweets and then attempts to fill in the gaps.

In order to attempt to validate the coverage of the Twapper Keeper server a comparison with the WTHashtag service's archive of the #www2010 hashtag showed that, on the same date, the WTHashtag service also reported that there were 3,616 tweets.

The potential loss of tweets and possible differences in the time taken by different harvesting services to harvest tweets will be documented in a Twapper Keeper FAQ to ensure that users wishing to publicise statistics on the numbers of tweets are aware of possible discrepancies in the figures provided by different services.

6.2. Use of Twapper Keeper APIs

The Summarizr service was developed independently of Twapper Keeper but made use of Twapper Keeper APIs. This service provides summaries and graphs of Twitter usage based on the Twapper Keeper data. This service, which was developed at Eduserv, an educational charity based in the UK, removes the need for developments

having to be provided by Twapper Keeper and vindicates the decision to encourage the take-up of the APIs by others. Independent discussions are taking place with the Summarizr developer on ways in which the Summarizr statistical summaries can themselves be reused by other applications.

As well as providing statistics on the total numbers of tweets and users for a hashtag as illustrated in Figure 2 the service also displays graphs showing the top Twitterers, @reply recipients, conversations, related hashtags and URLs tweeted [21].

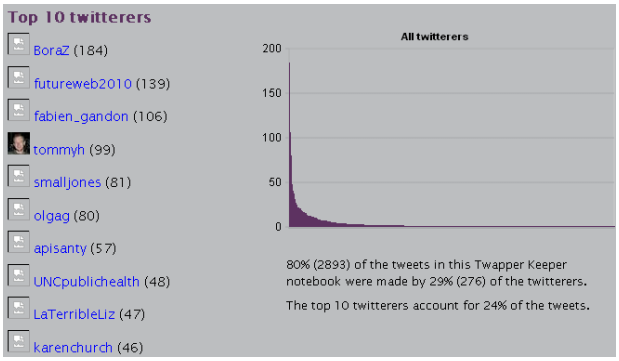

Figure 2: Use of Summarizr to display statistics of use **of** the for the #www2010 hashtag

The service also provides a display of a word cloud showing the relative frequency of the most popular words tweeted with a particular hashtag as illustrated in Figure 3.

Figure 3: Summarizr display of popular words

6.3. Summary of Status of Twapper Keeper Service Use

As of 1 July 2010 the Twapper Keeper archive contains 1,243 user archives, 1,263 keyword archives and 7,683 hashtag archives. There are a total of 321,351,085 tweets stored. The average number of tweets ingested per second is from 50 to 3,000 per minute (around 180,000 per hour. or 4.32 million per day). Since Twitter itself processes about 65 million tweets per day the Twapper Keeper service is currently processing about 6-7% of the total public traffic.

7. FURTHER DEVELOPMENTS

Recent developments to Twapper Keeper and Summarizr are storage and display of geo-location data. We invited participants at the IWMW 2010 event in July 2010 to geo-locate their tweets which enabled a map of the locations of Twitter users to be produced thus providing evidence of the remote participation at an event [12] as shown in Figure 4.

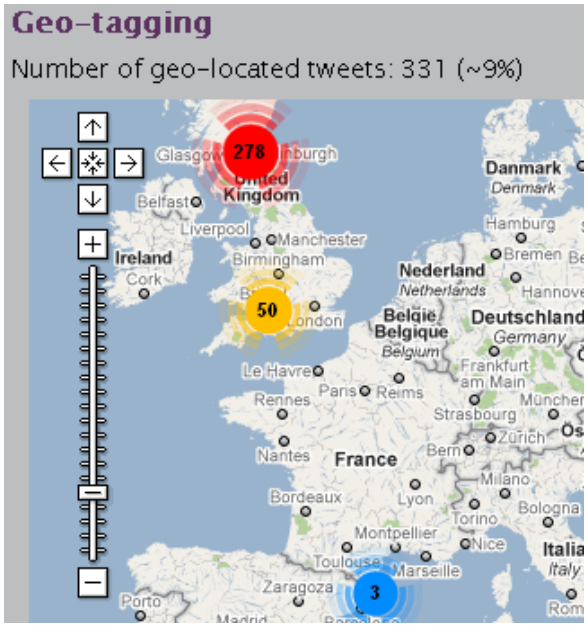

Figure 4: Display of geo-located tweets

In addition we used the Twapper Keeper APIs in conjunction with the Twitter captioning service to provide a captioned version of recordings of plenary talks shortly after the videos had been published.

We have identified the need to ensure that Twitter users are aware of the implications of Twitter archiving services. We will be developing guidelines which will help to raise awareness of the ways in which tweets could be reused, the possible risks which this may entail and approaches they can take to minimise such risks, including deletion of tweets from Twitter and archiving services which support deletion.

8. SUSTAINABILITY CHALLENGES

Although the JISC funding has been used to fund development work to address the needs of the UK higher education community and to support the migration of the service to a more stable platform this pump-priming funding cannot guarantee the sustainability of the service in the long-term.

The software developments which have been funded will be made available under an open source licence, thus allowing the Twapper Keeper service to be recreated if the host service were to disappear. In addition the data itself is available in a rich format allowing the data to be easily migrated to other environments.

The policy decision to fund development of a service provided by a commercial provider reflects the changing funding environment in the UK's public sector, in which the government has announced significant reductions in future investments in the sector.

The approaches taken in funding Twapper Keeper developments provides a useful experiment in alternative approaches to development work which will inform other development activities funded by the JISC.

9. CONCLUSIONS

This paper has described the importance of the archiving of Twitter posts by outlining case studies based on the ability to have reliable and consistent access to tweets. Rather than commissioning a new service the JISC has funded development of an existing 'cloud' service provided by Twapper Keeper. The approaches to ensuring the sustainability of this investment have been described.

The paper has summarised the requests received from the user community on developments to the services and reviewed the technical and policy challenges which the development work has faced.

The paper has described the experiences gained in use of the Twapper Keeper service to archive tweets from a large international conference and concluded by summarising developments which were deployed at a national event in the UK.

ACKNOLWEDGEMENTS

Acknowledgements are given to the JISC for their support for the Twapper Keeper development project.

REFERENCES

[1] Diakopoulos. N. and Shamma, D. *Characterizing Debate Performance via Aggregated Twitter Sentiment.* CHI 2010, ACM, 2010.

[2] Hawksey, M. *Twitter powered subtitles for BBC iPlayer*, MASHe blog, 16 Feb 2010, <http://www.rsc-ne-scotland.org.uk/mashe/2010/02/twitter-powered-subtitles-for-bbc-iplayer/> 2010.

[3] Hawksey, M. *Gordon Brown's Building Britain's Digital Future announcement with twitter subtitles*, MASHe blog, 23 Mar 2010, <http://www.rsc-ne-scotland.org.uk/mashe/2010/03/gordon-browns-building-britains-digital-future-announcement-with-twitter-subtitles/> 2010.

[4] Hawksey, M. *Presentation: Twitter for in-class voting and more for ESTICT SIG*, MASHe blog, 20 Apr 2010, <http://www.rsc-ne-scotland.org.uk/mashe/2010/04/presentation-twitter-for-in-class-voting-and-more-for-estict-sig/ > 2010.

[5] Hirst, A. *Twitter Gardening – Pruning Unwanted Followers*, OUseful blog, 24 Sep 2009, <http://blog.ouseful.info/2009/09/24/twitter-gardening-pruning-unwanted-followers/> 2009

[6] Hirst, A. *More Thinkses Around Twitter Hashtag Networks: #JISCRI*, <http://blog.ouseful.info/2009/09/04/more-thinkses-around-twitter-hashtag-networks-jiscri/> 2009.

[7] Hirst, A. *Twitter Powered Subtitles for Conference Audio/Videos on Youtube*, OUseful blog, 8 Mar 2009, <http://blog.ouseful.info/2009/03/08/twitter-powered-subtitles-for-conference-audiovideos-on-youtube/> 2009.

[8] JISC. *JISC Strategy* 2010-2012, <http://www.jisc.ac.uk/aboutus/strategy/strategy1012.aspx> 2010.

[9] Kelly, B. *Use of Twitter at the ALTC 2009 Conference*, UK Web Focus blog, 14 Sep 2009, <http://ukwebfocus.wordpress.com/2009/09/14/use-of-twitter-at-the-altc-2009-conference/> 2009.

[10] Kelly, B. *Privatisation and Centralisation Themes at JISC 10 Conference*, UK Web Focus blog, 15 Apr 2010, <http://ukwebfocus.wordpress.com/2010/04/15/privatisation-and-centralisation-themes-at-jisc-10-conference/> 2010.

[11] Kelly, B. *Evidence on Use of Twitter for Live Blogging*, UK Web Focus blog, 4 Aug 2009, <http://ukwebfocus.wordpress.com/2009/08/04/evidence-on-use-of-twitter-for-live-blogging/>, 2009.

[12] Kelly, B. *Geo-locating Your Event Tweets*, UK Web Focus blog, 6 July 2010, <http://ukwebfocus.wordpress.com/2010/07/06/geo-locating-your-event-tweets/> 2010.

[13] Library of Congress. *How Tweet It Is!: Library Acquires Entire Twitter Archive*, Library of Congress blog, 14 Apr 2010 <http://blogs.loc.gov/loc/2010/04/how-tweet-it-is-library-acquires-entire-twitter-archive/> 2010.

[14] McGill, K. *Summary: What Is The Web?*, IWMW 2009 blog, 30 Jul 2009, <http://iwmw2009.wordpress.com/2009/07/30/summary-what-is-the-web/> 2009.

[15] Twitter. *Twitter Search API Method: search*, <http://apiwiki.twitter.com/Twitter-Search-API-Method:+search> 2010

[16] Twitter. *Terms of Service*, <http://twitter.com/tos> 2010.

[17] Twapper Keeper. *JISC-Funded Developments To Twapper Keeper*, Twapper Keeper blog, 16 Apr 2010, <http://twapperkeeper.wordpress.com/2010/04/16/jisc-funded-developments-to-twapper-keeper/> 2010.

[18] Twapper Keeper. *Study of Missed Tweets*, Twapper Keeper blog, 5 May 2010, <http://twapperkeeper.wordpress.com/2010/05/05/study-of-missed-tweets/> 2010.

[19] Twapper Keeper. *Hashtag notebook #jiscri tag for JISC Rapid Innovation projects which are creating software for Higher Education*, <http://www.twapperkeeper.com/hashtag/jiscri> 2010.

[20] Sakaki, T., Okazaki, M., and Matsuo, Y. *Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors*, International World Wide Web Conference Proceedings of the 19th international conference on World wide web, Raleigh, North Carolina, USA. <http://portal.acm.org/citation.cfm?id=1772690.1772777> 2010.

[21] Summarizr. *TwapperKeeper Archive for hashtag notebook www2010*, <http://summarizr.labs.eduserv.org.uk/?hashtag=www2010> 2010.